OSU Working Papers in Linguistics 55, 23-42

The Interplay of Perception and Phonology in Tone 3 Sandhi in Chinese Putonghua*

Huang, Tsan
huang@ling.ohio-state.edu

1. Introduction

The phenomenon of tone sandhi has long been noticed in the Chinese dialects (e.g. Chao 1948, 1968). Past studies allude that tone sandhis may be analyzed as processes leading to ease of articulation (Chao 1948, 1968; Cheng 1973; among others). But those analyses do not offer an explanation why one particular output is selected when other outputs are possible. Take for example the Tone 3 (or T3) sandhi process of standard mainland Mandarin Chinese, which is examined in this study. While simplification seems to be the correct analysis in that both native intuition and vocal physiology support the theory that it is hard to produce two dipping T3s in a row[1], it does not explain why T3 simplifies to T2 but not T1 or T4, the other two "simpler" tones in Mandarin Chinese. (See Section 2.1 for a detailed description of the tones in Standard Mandarin Chinese.)

In the present study, we hypothesized that the output of the third tone sandhi may be perceptually conditioned. In the words of Kohler (1990), Hura et al. (1992), and

* I would like to express my gratitude to my teachers Keith Johnson, Beth Hume, Michael Broe, and Dave Odden, and to members of the PiP seminar, Prof. San Duanmu at University of Michigan, Janice Fon, Martin Jansche, Georgios Tserdanelis, and all others in the OSU Linguistics Department who offered their help and support. Needless to say, all errors are mine.

[1] T3 surfaces as a low (falling) tone in most cases (see Section 2.1). The trigger of T3 sandhi seems to be the "lowness" of T3 (see also Shih 1997). C. C. Cheng (1968, cited in Shih 1997) reported that, even when Mandarin speakers code-switch between Chinese and English, a T3 would change into T2 when the first syllable of the following English word is unstressed – i.e. bearing a low tone:

/hao^{214} pro'fessor/ → [hao^{35} professor] "good professor"; but
/hao^{214} 'student/ → [hao^{21} student] "good student".

Steriade (2001), this may be a case of perceptually tolerated articulatory simplification. That is, T3 is selected as the sandhi form because it is perceptually more similar to T3 than T1 or T4 is, which makes the change relatively hard to detect in perception. Although there are quite a few studies on T3 sandhi and on the confusability of T3 and T2 in the literature, none of these studies compared the perceptual confusability of T3 and T2 with that of T3 and T1, nor with that of T3 and T4. Thus, none of them dealt with this issue directly. The present study tries to address this gap in the literature with a perceptual experiment of monosyllabic tonal pairs, which recorded both the "same"/"different" judgements made by the participants and their reaction time during response latency. We hope that the results of this study will help provide better understanding of this sandhi process and, more importantly, some insight into the interplay of perception and phonology (Hume & Johnson, 2001).

2. Background

2.1 The tone sandhi phenomenon in Standard Mandarin Chinese (or Putonghua)

In almost all Chinese dialects, underlying full tones may be modified under the influence of their tonal phonetic environment. This phenomenon is known as tone sandhi (see, for example, Chao 1948, Kratochvil, 1968). In this study, we looked at the tone sandhi phenomenon in standard mainland Mandarin Chinese, or Putonghua. This language has four lexical tones: level high [55, ˥], mid-rising [35, ˧˥], low falling-rising [214, ˨˩˧][2], and high falling [51, ˥˩], traditionally termed Tones 1, 2, 3, and 4, respectively. (The numbers in the square brackets indicate the pitch values of these tones on a five-level scale. And the drawings next to the numbers are graphic representations of those pitch values, termed Chao's tone letters; for a detailed discussion of the tone letter system, see Chao 1948 & 1968.) There is also a "fifth" tone, namely the inherent neutral tone, whose pitch value varies dependent on its preceding full tone.

As described in Chao (1948, 1968), the third-tone sandhi happens when T3 of Chinese Putonghua (– valued 214) becomes T2 (– valued 35) when immediately followed by another T3. Schematically, / ˨˩˧ ˨˩˧ /→ [˧˥ ˨˩˧].

It is claimed by many that morphological and syntactic boundaries are irrelevant here. Some examples are provided in (1) below:

(1) a. /hao^{214} mi^{214}/ → [hao^{35} mi^{214}] "good rice"
modifier head noun

b. /mi^{214} hao^{214}/ → [mi^{35} hao^{214}] "The rice is good."
subject predicate

[2] In Cheng (1973), T3 is described as having the value [315].

c. /mai^{214} mi^{214}/ → [mai^{35} mi^{214}] "to buy rice"
| |
verb object

Other phonetic variants of T3 include [21][3] and [214], the first of which appears before all full tones except T3 and the second of which appears in sentence-final position.

2.2 Perception and phonology

Phonologists have noticed the influence of perception on phonology from very early on (Trubetzkoy, 1969). In Jakobson, Fant, and Halle (1952), perceptual features are treated as primary (see also Jakobson and Halle, 1956). But the generative tradition of phonology since Chomsky & Halle (1968) centers around articulatory phonology. Now a revival of the view of the interplay of perception and phonology seems to be in process. People have been asking questions such as "To what extent do speech perception phenomena influence phonological system?" "To what extent does the phonological structure of language influence speech perception?" (Hume and Johnson, 2001).

Kohler (1990), Hura et al. (1992), and Steriade (2001) hold that phonological processes such as segmental reduction, deletion, and assimilation are perceptually tolerated articulatory simplification and that the direction of such processes is determined by perception. That is, these processes only take place when the output of such a change is found to be highly confusable with the input perceptually. For example, as place contrasts in sibilants are more salient than place contrasts in stops (Kohler 1990, Hura et al. 1992), the following patterns were found in the common retroflexization process in Sanskrit (Steriade 2001):

(2) Sanskrit apical assimilation in $VC_1C_2(C_3)V$ sequence

a. same manner apical clusters: progressive assimilation
/V**ʈ**tV/ → [V**ʈʈ**V]
b. sibilant-stop clusters: progressive assimilation
/V**ʂ**tV/ → [V**ʂʈ**V]
c. stop-sibilant clusters: no assimilation
/V**ʈ**stV/ → [V**ʈ**stV]

In (2a) and (2b), the underlying dental stop /t/ surfaces as the retroflex [ʈ] as a result of assimilation to the place feature of the proceeding /ʈ/ or /ʂ/. In (2c), however, as place contrast between the dental /s/ and the retroflex /ʂ/ is more salient than the place contrast between /t/ and /ʈ/ -- thus, /s/ and /ʂ/ are less confusable than /t/ and /ʈ/ are -- assimilation does not happen and /s/ surfaces as [s].

[3] Some writers treat [21] as the underlying form of T3, as this is the most common surface shape. In fact, in the variety of Mandarin spoken in Taiwan, [21] surfaces in the final position. Sometimes, it may even surface in the final position in Putonghua.

If perception can influence segmental phonology in such a way, it seems reasonable to hypothesize that it may also have an impact on suprasegmental phonology and that it may play a role in the T3 sandhi process in Chinese Putonghua. Kiriloff (1969) found that syllables with Tone 2 and Tone 3 may be perceptually confusable. Fon (1997, MA thesis) and Fon et al. (1999 ms.) observed that both T2 and T3 have a dip in them (see also sources cited therein, e.g. Ho 1976, Yang 1995). Although the initial dip in T2 is usually ignored in phonological analysis, it has been shown to be perceptually important (see, for example, Gottfried & Suiter 1997). In a binary forced choice (T2 or T3) experiment where subjects were asked to label the tones whose pitch contours had been manipulated, Shen & Lin (1991) found that both the intrinsic duration of these tones and the turning points (i.e., where the rise starts in the contour) contribute to the confusability (see also Blicher et al. 1990, Chuang et al 1971, and Fon et al. 1999, ms.). Unfortunately, none of these studies can be used as convincing evidence to support our hypothesis that T3 sandhi is perceptually conditioned, as no comparison was made between the degree of confusability of T3 and T2, and that of any other tonal pairs in this language. In addition, we were also interested in finding out how phonology may influence listeners' perception of tones. Thus, the following experiment was designed to test our hypothesis directly.[4]

3. The experiment

3.1 Participants

Ten Chinese listeners (6 female, 4 male, average age 27.9) and thirteen American English listeners (7 female, 6 male, average age 21.8) were recruited from the Columbus campus of the Ohio State University (OSU). The Chinese listeners were graduate students (or their spouses) at OSU. Although a couple of them are not from the geographical regions where Mandarin is spoken, they were all fluent in the standard language due to their education background: they all received at least college education, and Putonghua is usually the language of instruction in most classrooms in mainland China. The English listeners were undergraduate students taking an introductory linguistics course at OSU. They were all native speakers of Ohio English. The Chinese were paid for their participation in the experiment, whereas the Americans earned extra credit points for their Linguistics 201 class.

[4] There is historical evidence that the T3 sandhi may have happened 700 years ago when the T3 contour could have been completely different from its current shape. This process may have been grammaticized in the Mandarin dialects and carried down to the present day. But it is not the case that all current Mandarin dialects preserve this sandhi rule. For example, it is no longer in my dialect, Rugaohua, a Jianghuai Mandarin dialect. We may speculate that a certain generation of Rugaohua speakers gave a second thought of the sandhi process and, due to a change in the tonal shape of T3, could not see why it was necessary to have the process. So, they decided to drop it. On the same basis, maybe speakers of Beijing Mandarin, the base language for Putonghua, did a similar analysis. But since it was still necessary to have the sandhi, it was reinvented. And the fact that Putonghua speakers apply the rule even when code-switching (see Footnote 1) is evidence that it is a productive synchronic phonological process. Thus, a synchronic analysis seems to be justified.

The American English listeners were included to see if T3 and T2 of Chinese Putonghua in the T3 sandhi environment share some property that makes them confusable to non-native listeners. We assume that, if there is no effect of the listener's native phonology on perception, phonetic universality should allow everybody to act the same. Previous studies have shown that it is feasible to include "non-native" listeners. Kiriloff (1969) found that, when asked to ignore the segmental part of the syllable and focus on the tones, non-native speakers' performance was quite good (an average of 17.5 correct identifications out of 20 stimuli, or 87.5%)[5]. On the other hand, if we do find a difference between native and non-native listeners' performance, it might help us gain insight into how phonology may influence perception.

3.2 Stimuli

The stimuli were constructed from recordings produced by a female Putonghua speaker in disyllabic nonsense sequences with 15 tonal combinations (– that is, all possible pairs except T3T3 which does not occur in natural speech. The segmental make-up of these recorded sequences was kept constant and was always /bao-fang/. The typical stress pattern[6] for a disyllabic full-toned sequence was used to get the appropriate pitch contours of the four tones in the environment where T3 sandhi occurs. Ten (10) randomized lists of these sequences were recorded. The original recordings were done in a sound-proof booth in the phonetics laboratory at the OSU Linguistics Department. The speaker read from the afore-mentioned 10 randomized lists and was recorded with a head-mounted microphone (Shure SM10A model) and a DAT recorder.

The recordings were digitized at 22,050Hz with 16 bit samples. The first syllable (i.e. /bao/) was cut from these sequences. The seven (7) best productions of these /bao/ syllables for each of the four tones (as determined subjectively by the author) were then chosen to splice the test stimulus pairs, while three (3) other productions were used in the training session.

Figures 1, 2, 3 and 4 show pitch tracks of these stimulus tones. Note that only the first "half" of the T3 tonal contour is realized, which is typical of T3 in this non-final position.

---

[5] Gottfried & Suiter (1997) did find some degree of performance difference in native versus non-native listeners. But it was a more difficult identification task and their four conditions (– namely, initial only, center only, silent center, and final only) all involved cutting off some part of the syllable. Lee et al. (1996) also found a small native speaker advantage.

[6] Yip (1980) and Zhang (1988) mentioned that T3 sandhi is conditioned by the metrical pattern of the utterance and that the T3 that undergoes the sandhi has to be in the weak branch of the stress matrix, i.e. the syllable bearing the sandhi tone must not be linked to a node at the highest/primary stress level. Thus, it is predicted that the first T3 in /xiao3jie3/ 'miss' (with a weak-strong pattern) would undergo the T3 sandhi and surface as [xiao2jie3], whereas that in /jie3jie3/ 'older sister' (with a strong-weak pattern) would not, yielding the surface form [jie3jie0]. But see Shih (1997), where she holds that stress does not play a role in the sandhi processes. We chose not to commit ourselves to any particular phonological framework here and tried to take into consideration all possible conditions for this sandhi process.

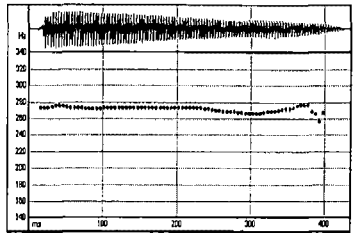

Figure 1. Pitch track of stimulus T1

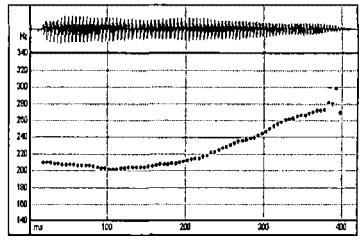

Figure 2. Pitch track of stimulus T2

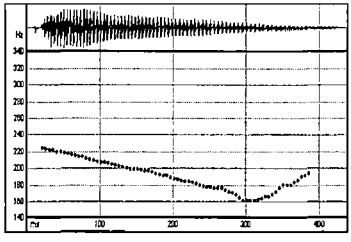

Figure 3. Pitch track of stimulus T3

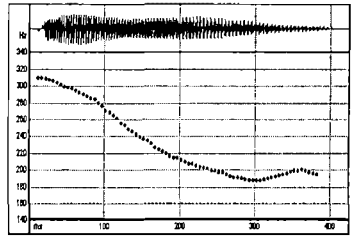

Figure 4. Pitch track of stimulus T4

The test session consisted of 7 sections, each of which contained 20 stimulus pairs. Thus, all participants listened to 20 × 7 = 140 pairs of the form /bao-bao/. The 20 pairs in each section included 12 different pairs (see the checked boxes, marked with x, in Table 1 below) and 8 identical pairs (i.e., each of the 4 identical pairs in the empty boxes in Table 1 was repeated twice in any of the test sections)[7]. Only the results of different pairs were analyzed. The identical pairs were included as fillers.

Table 1. Tonal combinations to be tested

	T1	T2	T3	T4
T1		x	x	x
T2	x		x	x
T3	x	x		x
T4	x	x	x	

[7] Each identical pair contains two repetitions of the same .wav file. Thus, the experiment emphasized psychoacoustic discriminability of tones.

3.3 Method

A discrimination task was used. Participants were tested in front of a computer one at a time in a sound-proof booth. The stimuli were presented to them through headphones, using the Micro Experimental Lab (MEL) program installed on a PC. While each stimulus pair was played with a 2000ms inter-pair interval, the words "same" and "different" were also displayed visually on the left and right sides of the computer screen, respectively. The participant input responses by pressing the "same" or "different" buttons on a button-box connected to the PC. Participants were asked to use their left and right index fingers to press the "same" and "different" buttons, respectively. Instructions, both given orally by the experimenter during the training session and displayed visually on the PC screen during the test session, asked the participant to respond as accurately and as quickly as possible. After each correct "same"/"different" judgment was made, the reaction time (RT) would appear on the screen as feedback; otherwise, the screen would display the words "wrong response". This made it clear to the subjects what a good performance was: one with shorter reaction time and fewer errors.

Both the "same-different" judgement accuracy and RT were recorded as experiment results. The measurement for RT started from the onset of the second syllable of the stimulus pair.[8]

3.4 Predictions

We predicted that, if T2 and T3 are more confusable, i.e. closer to each other in the perceptual space, then (i) people would make more mistakes when asked to tell whether they are the same or different, and (ii) people would take longer to make the judgment, that is, the shorter the perceptual distance, the longer the reaction time (RT) (see, for example, Shepard et al. 1975, Shepard 1978, Takane et al. 1983, Nosofsky 1992, although these authors disagree on what exactly the relationship between perceptual distance and RT is and how the transformation between them should be done. We shall postpone the discussion on these issues until Section 5.)

4. Results

The results basically support our hypothesis that T2 and T3 are perceptually more confusable. In terms of the mistakes that listeners made, there was no statistically significant difference between the tonal pairs, as error rates were very low in the responses of both the Chinese and English groups. But the pairs T2-T3 and T3-T2 did attract more errors than other pairs. Table 2 shows mean RT values of correct "different" responses and error rate in percentage for each non-identical tonal pair.[9]

---

[8] The mean duration measurements for all stimulus syllables are: T1=375.9ms, T2=414ms, T3=389.5ms., and T4=387.8ms. Such differences do not seem to be big enough to affect the RT measurements, as the adjusted RT data (with the duration of the second syllable subtracted) show a similar pattern.

[9] The median RT data– with or without the duration of the second syllable – reveal a pattern similar to the mean RT data (see Appendix I).

Table 2. Mean RT (ms) for correct "different" responses and percentage of errors

TONAL PAIRS	T1/T2		T1/T3		T1/T4	
	T1T2	T2T1	T1T3	T3T1	T1T4	T4T1
Chinese	568.9(4%)	556.7(7%)	572.8(3%)	584.2(6%)	602.4(4%)	572.6(4%)
English	558.1(1%)	671.5(1%)	516.6(2%)	556.1(2%)	606.8(5%)	594.0(2%)
TONAL PAIRS	T2/T3		T2/T4		T3/T4	
	T2T3	T3T2	T2T4	T4T2	T3T4	T4T3
Chinese	699.4(11%)	667.4(7%)	512.1(0%)	583.2(4%)	542.9(0%)	547.0(4%)
English	748.4(16%)	663.5(13%)	615.1(11%)	568.6(3%)	591.0(5%)	624.3(2%)

We can see that the Chinese listeners scored 62 correct responses out of all 70 T2T3 stimulus pairs (= 7 sections × 10 participants) with an error rate of 11% and 65 correct responses out of all 70 T3T2 stimuli with an error rate of 7%. On the other hand, the English listeners scored 76 correct responses out of all 91 T2T3 stimulus pairs (7 sections ×13 participants) with an error rate of 16% and 79 correct responses out of all 91 T3T2 stimulus pairs with an error rate of 13%.

Although error rates were too low to be significant, the RT data turn out to be very informative. The graphic representation in Figure 5 may help us see clearly what the RT values for the T2/T3 pairs are like compared to other tonal pairs. The points on the X-axis represent the non-identical pair types, and the numbers along the Y-axis show reaction time in milliseconds. The solid line represents the Chinese listeners' data, while the dashed line the English listeners'.

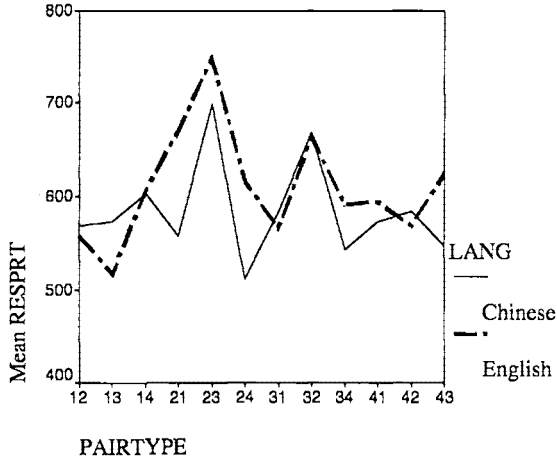

Figure 5. Mean RTs (in milliseconds) for the correct responses

As we expected, the slowest RT for the Chinese participants was found with the T2/T3 pairs (the two peaks in the solid line in Figure 5), with the RT for T2T3 being even longer than that for T3T2. One possible explanation for this pattern might be that, as this sequence is identical to the sandhi output where the T3 and T2 distinction is neutralized into T2 for the Chinese listeners, they are biased because of their native phonology. However, we see a similar picture with the American listeners: the RT for T2T3 is also the longest, and the RT for T3T2 is the third longest of all pairs (shorter than that for T2T1), which may be seen as evidence for saying that the way the Chinese reacted is not completely due to their native phonology: phonetically, there exists some universal perceptual distance between these tones for both the native and non-native listeners.

5. Analyses

5.1 Repeated measures analysis of variance and Independent-Samples T test

A repeated measures analysis of variance (ANOVA) was performed on the RT data of the listeners' correct "different" responses, with all 12 non-identical tonal pairs (i.e. T1T2, T2T1, T1T3, T3T1, T1T4, T4T1, T2T3, T3T2, T2T4, T4T2, T3T4, T4T3) as within subject variables, and language as between subject variable. No significant result was found between listener/langauge groups. But there was a significant effect with pair types, sig.[$F(9.321, 1118.461) = 353343.626$, $p < .001$, $\tilde{\eta}^2 = .106$]. There was also a significant effect with the interaction of language and pair, sig.[$F(9.321, 1118.461) = 103801.579$, $p < .001$, $\tilde{\eta}^2 = .034$].

A post-hoc test of pair-wise comparison, which compares the raw RT values for each pair against all other pairs within the same listener group, shows that pairs T2T3 and T3T2 are significantly different from the other pairs ($p<.05$) for both groups of listeners. For the Chinese listeners, pairs T2T3 and T3T2 are totally different things from the other pairs($p<.05$). Pair T2T3 was found to be significantly different from all pairs except pairs T1T4 and T3T2. Pair T3T2 is significantly different from all pairs except pairs T4T1, T4T2, T3T1, T1T4 and T2T3, showing a possible effect of phonology on perception, as no difference was found between any two of the other pairs. Interestingly, pair T2T3 was found to be significantly different from three more pairs than pair T3T2, which seems to make the effect of phonology even stronger, as T2T3 is the output of the T3 sandhi.

The English listeners, on the other hand, found pair T1T3 to be the least confusable and significantly different from all other pairs except T3T1 and T1T2 ($p<.05$). They also found pairs T2T3, T2T1 and T3T2 to be the most confusable ($p<.05$). This pattern seems to be more phonetic than phonological, as the English listeners seem to rely more on the phonetic shapes of these tones when making their decisions. If the pitch value of the ending point of the first syllable matches that of the starting point of the second syllable, the English listeners found them to be confusable. This behavior is different from that of the Chinese listeners' who found only the T2 and T3 pairs to be confusable. Again, this seems to provide more evidence that the Chinese listeners' perception is influenced by their native phonology.

An Independent-Samples T test was also performed on the RT data, with language as the grouping variable and RT as the test variable. The outliers in both listener groups were taken out. The results are as follows:

Table 3. Independent-Samples T test

Pair type		*meanRT Chinese*	*meanRT English*
T1T2	t(137) = 1.175, p = .242	530.9	508.5
***T2T1**	**t(133) = -4.677, p < .001**	**514.1**	**608.6**
***T1T3**	**t(138) = 2.41, p = .017**	**543.5**	**496.1**
T3T1	t(138) =.665, p = .507	550.8	536.4
T1T4	t(138) = -1.784, p = .077	554.5	591.8
T4T1	t(138) = -1.359, p = .176	543.8	572.6
T2T3	t(120) = -1.022, p = .309	663.3	687.2
T3T2	t(135) = .345, p = .731	662.7	654.4
***T2T4**	**t(128) = -4.667, p < .001**	**486.8**	**578.3**
T4T2	t(138) = 1.098, p = .274	575.6	551.4
***T3T4**	**t(141) = -3.02, p = .003**	**513.5**	**579.7**
T4T3	t(135) = -1.688, p = .094	532.9	568.9

The pairs with a significant between-language-group effect ($p<.05$) have been bold-faced and indicated with an asterisk in Table 3. The mean RT values (in milliseconds) make the pattern even more interesting. In general, the Chinese listeners did better than the English listeners. For some pairs that the English listeners found confusable, i.e., **T2T1**, **T2T4**, and **T3T4**, the Chinese listeners did not seem to have more difficulty distinguishing them at all. For pair **T1T3**, which the English listeners found to be the least confusable, the Chinese listeners did not seem to see it as an easier pair than the other pairs.

5.2 Multidimensional scaling (MDS)

In a sense, the RT data obtained reflect similarity between the tones: RT values increase as the tones get more similar. We need to find a way to transform RT (or similarity) values into perceptual distances (or dissimilarity). As no direct measurements can be made for either the physical or the perceived distances between these tones, RT measurements were converted into perceptual distances based on the assumption that the closer two "objects" are in the perceptual space, the longer it takes for people to tell them apart (see, for example, Shepard et al. 1975, Shepard 1978, Takane et al. 1983, Nosofsky 1992).

How exactly RT reflects perceptual or physical distance is still a question begging to be answered. In our case here, we would probably also need to take into account the influence of phonology on perception as well as the characteristics of the stimuli (i.e., the phonetic characteristics of the tones). Nevertheless, several approaches have been proposed to convert RT into distances. Curtis et al. (1973), Shepard et al. (1975), and

Shepard (1978) advocate for the reciprocal function. Their argument for that is, with correct "different" judgments, reaction time values have been found to be nearly reciprocal of distance values. Takane and Sergent (1983) and Nosofsky (1992) suggest the log normal function. Takane and Sergent's (1983) reason for choosing the log normal function over the reciprocal function is that it is not the case that correct "same" RT is reciprocal to distance. As only the RTs of correct "different" judgments were of interest in the present study, the choice of the reciprocal approach seems to be justified. In addition, this approach is well-supported by previous research.

In fact, we did make use of the log normal approach and found the MDS results to be very similar to the reciprocal approach. In addition to the reciprocal and the log normal functions, which turn linearly related RTs into a non-linear distribution of distances, we also tried a linear approach suggested by Michael Broe (personal communication). RTs were rescaled using the formula (Observed RT/Maximal RT) so that they now distribute along a scale of 0~1. Then, the 0~1 RT scale were turned into a 0~1 distance scale by subtracting the new "RT" values from 1 (i.e. distance = 1 - Observed RT/Maximal RT]). Again, the MDS results are surprisingly similar to the reciprocal approach. The calculated distances (=1/RT) are reported in Table 3 below.[10]

Table 4. Distances derived from RTs for correct responses for all different tonal pairs. Values are 10^3 times the original reciprocal values.

	T1T2	T1T3	T1T4	T2T1	T2T3	T2T4
Chinese	1.895	1.903	1.853	1.981	1.54	2.121
English	1.977	2.079	1.766	1.617	1.468	1.778
	T3T1	T3T2	T3T4	T4T1	T4T2	T4T3
Chinese	1.871	1.651	2.014	1.905	1.874	1.95
English	1.918	1.570	1.860	1.800	1.884	1.800

The mean distance for tonal pairs involving the same tones was taken to be the distance between these tones in the MDS analysis. For example, the mean value for the T1T2 and T2T1 distances was taken to be the distance between T1 and T2.

Table 5. Averaged distances for tonal pairs involving the same tones.

	T1/T2	T1/T3	T1/T4	T2/T3	T2/T4	T3/T4
Chinese	1.938	1.887	1.879	**1.596**	1.998	1.982
English	1.797	1.998	1.783	**1.519**	1.831	1.830

As shown in Table 4, the distance between T2 and T3 was found to be the shortest by both the Chinese-speaking and the English-speaking participants.

[10] Distances are averages of all reciprocal values of the original RT data, not the mean RT.

These data were then analyzed as dissimilarities using monotonic MDS, with the 2-dimensional MDS taken as the default.[11] And evaluation of a 2-dimensional scaling is very satisfactory: stress for the Chinese listeners' data is 0.0 and that for the English listeners' data 0.00174; values of the proportion of variance (RSQ) are 1.0 and 0.99 for these data sets, respectively[12]. A cluster tree analysis for both the Chinese and the English (Figure 8) listeners' data revealed that the groupings of the tones are exactly the same for both groups of listeners: T3 is grouped with T2, and T4 is grouped with T1.

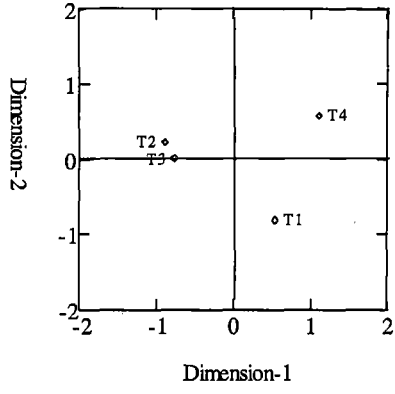

Figure 6. Two-dimensional scaling for Chinese listeners' RT data. [stress = 0.00, Proportion of variance (RSQ) = 1.0]

[11] The error data in percentage (see Table 2) was also analyzed as similarity data; that is, it was assumed that the more similar the two objects are, the higher the error rate is. The MDS results turned out to be very similar to the distance analysis (see Appendix II).

[12] As we only have four (4) objects in the analysis, the stress is low in a 1-dimensional analysis, too. But it does show some improvement for the English listeners' data when we change the number of dimensions from 1 to 2.

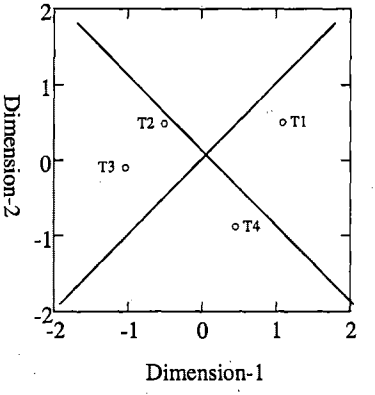

Figure 7. Two-dimensional scaling for the English listeners' RT data. [stress = 0.001729, Proportion of variance (RSQ) = 0.99902]

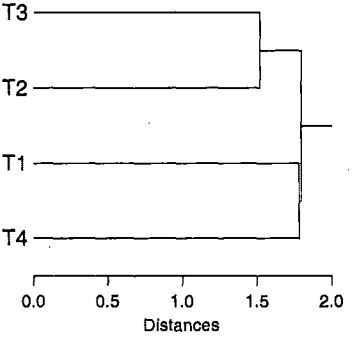

Figure 8. Groupings of the four Putonghua tones in Chinese/English data.

6. Discussion

Combining the information of the distances between the tones along the two dimensions in Figures 6 and 7 and the information of their groupings in Figure 8, we have a very telling picture. First, notice that T2 and T3 are grouped together in both the Chinese and the English listeners' data. This pattern, consistent with the pattern shown in Figure 5, shows that these two tones share some intrinsic phonetic property that can be perceived by both native and non-native speakers of Chinese Putonghua. This may be seen as evidence for our hypothesis that the sandhi process is allowed to take place because such a change is relatively hard to detect and that the selection of T2 as the output of the sandhi may be perceptually conditioned.

Second, the perceived distance between T2 and T3 seems to be smaller relative to the other tonal pairs for the Chinese listeners than the inter-pair difference for the English listeners.[13] Recall that the ANOVA and post-hoc test also show that the Chinese listeners treated the T2/T3 pairs as being different from all other tonal pairs. These findings provide evidence for the view that speech perception is influenced by phonology (see Hume and Johnson 2001). That is, because of the influence of the phonological structure of their native language, in this case the T3 sandhi rule, the Chinese listeners were highly biased toward the similarity of pairs T2T3 and T3T2. On the other hand, the English listeners were dealing with mostly the phonetic characteristics of the tones. (Basically, as was mentioned before, if the ending pitch of the first syllable matches the starting pitch value of the second syllable, for example T2T1, or if two tones in a pair share a similar tonal contour, for example T3T4, the pair was found to be more confusable.) If there is no phonological effect in addition to familiarity, the MDS for the Chinese listeners would look different: one would expect the Chinese listeners' perceived distance between any tonal pair to be longer than the English listeners' due to familiarity. The distance between T2 and T3 might still be short for the Chinese listeners relative to the other tonal pairs because of the intrinsic properties of these tones. But the overall MDS pattern should look similar to the pattern that we saw in the English listeners' data but with a longer distance between T2 and T3 as compared to that in the English listeners' MDS.

It may not be very obvious what the two dimensions in the MDS configuration are, especially in the English listeners' data. The added (diagonal) lines in the configuration figures (Figures 6 & 7), which try to capture the information given in the cluster trees, may be seen as (rotated) axes. These (rotated) axes show that, along one dimension, both the Chinese-speaking and the English-speaking listeners have divided the tones into two register ranges according to the F0 values at the beginning of the tones. (For the pitch tracks, refer to Figures 1 through 4 in Section 3.2.) Thus, T2 and T3, both of which start with a F0 value that falls in the middle of the speaker's pitch range, are grouped together. And so are T1 and T4, both of which start with a F0 value that falls in the upper level of the speaker's pitch range. Along the other dimension, the tones seem to have been grouped together according to the characteristics of their pitch contours. Thus,

[13] The absolute T2/T3 distance value for the Chinese listeners is longer than that of the English listeners, which may be another effect of phonology on perception, as they perceive tones better than the English listeners who speak a non-tone language.

T1 is set apart from T2, T3 and T4 because T1 is a level tone with a static pitch level while the other three tones are contour tones with dynamic pitch movements. In other words, MDS reveals that both the tonal contour and the starting pitch point are important cues for tonal perception.

The patterns in the result of the Independent-Samples T test are also very revealing. It provides evidence for the view that the Chinese listeners treat each tonal contour as an indivisible unit (see also Jansche 1999 ms.), as neither the phonetic pitch level of the starting or ending point of the contour nor the similarity in tonal contours seems to contribute much to the confusability or distinctiveness of the tones. Unlike the English listeners who were using these characteristics of the tones as important phonetic cues to distinguish the tones, the perception of the Chinese listeners seemed to be independent of these cues to a certain extent. In other words, the Chinese listeners' phonological knowledge seems to have "transcended" their phonetic knowledge. The fact that, in one case, with pair **T1T3**, the Chinese even "suffered" from their phonological knowledge – i.e., failed to use the phonetic cues as effectively as the English listeners did – also shows that tonal perception is not influenced merely by familiarity; otherwise, one should expect the Chinese listeners to do better in all cases.[14] They did not. As can be seen in Table 3, they treated pair **T1T3** as an "average" pair. They performed almost as poorly as the English listeners did on pairs T2T3 and T3T2. We are not denying that familiarity played a role here, as the Chinese did better in general. Familiarity might have interacted with phonology, as the Chinese listeners did perceive the T2T3 and T3T2 pairs slightly better than the English listeners did, although, given their native phonology, one might not have been totally surprised should the Chinese have appeared to be "blind" to the distinction between T2 and T3. If one takes a second look at the error data shown in Table 2, he may find a similar pattern there. That is, the mistakes that the English listeners made were more phonetically-driven, while the mistakes in the Chinese listeners' data point to the influence from the Chinese tonal phonology.

7. Conclusion

To sum up, we examined Chinese Putonghua tones with a "same"/"different" discrimination task in this study. Distances between the tones were derived from the reaction time data by the reciprocal function. The MDS analysis on both the RT data and the error data shows that T3 is perceived as closer to T2 than it is to T1 or T4, which supports our hypothesis that T2 is chosen as the sandhi form for T3 because such a change is perceptually tolerated. The MDS analysis also shows that phonology influences speech perception, as the Chinese listeners perceived an even shorter relative distance between T3 and T2 than the American English listeners did. This is further supported by the post-hoc test result which shows that the Chinese listeners found only the T2/T3 pairs to be significantly different from all other tonal pairs. The Independent-Samples T test also reveals a phonological effect on perception as the Chinese appeared to have treated

[14] In two experiments involving more complicated tasks of tonal discrimination, Lee et al. (1996) found that native tone language speakers did better than speakers of a different tonal language, who, in turn, did better than nontone language speakers.

each tonal contour as an indivisible unit and ignored some important phonetic cues. It is evident from these results that there clearly is an interplay between perception and phonology and that the two may interact to constrain changes in the phonological structure of the language.

REFERENCES

Blicher, D. L., Diehl, R. L., & L. B. Cohen, 1990. Effects of syllable duration on the perception of the Mandarin Tone 2/Tone 3 distinction: evidence of auditory enhancement. *Journal of Phonetics*, 18 (1), 37-49.

Chao, Yuen Ren. 1948. *Mandarin Primer.* Cambridge, MA: Harvard University Press.

----------. 1968. *A Grammar of Spoken Chinese.* Berkeley: University of California Press.

Chen, Matthew Y. 2000. *Tone Sandi patterns across Chinese dialects.* Cambridge, UK: Cambridge University Press.

Cheng, Chin-chuan. 1973. *A Synchronic Phonology of Mandarin Chinese.* The Hague: Mouton.

Chomsky, N. & M. Halle, 1968. *The sound pattern of English.* New York, Harper & Row.

Chuang, C.K., S. Hiki, T. Sone, & T. Nimura, 1971. The acoustic features and perceptual cues of the four tones of standard colloquial Chinese. 7th International Congress on Acoustics, Budapest 1971.

Curtis, D. W., Paulos, M. A., & S. J. Rule, 1973. Relation between disjunctive reaction time and stimulus difference. *Journal of Experimental Psychology.*

Fon, Y-J. 1997. What are tones really like? – An acoustic-based study of Taiwan Mandarin tones. Master's thesis. National Taiwan University.

Fon, J., Chiang W.-Y., & H. Cheung, 1999. Production and Perception of T2 and T3 in Taiwan Mandarin. Ms.

Gottfried, T. L. & T. L. Suiter, 1997. Effect of linguistic experience on the identification of Mandarin Chinese vowels and tones. *Journal of Phonetics* 25: 207-231.

Hume, E. & K. Johnson, in press. A Model of the Interplay of Speech Perception and Phonology. In E. Hume & K. Johnson (eds.) *The Role of Speech Perception in Phonology* (pp 3-26). New York: Academic Press.

Hura, S. L., B. Lindblom, & R. L. Diehl, 1992. On the role of perception in shaping phonological assimilation rules. *Language and Speech*, 35(1,2): 59-72.

Jakobson, R., Fant, G., and M. Halle, 1952. *Preliminaries to speech analysis*: the distinctive features and their correlates. Cambridge: Acoustics Laboratory, Massachusetts Institute of Technology.

Jakobson, R. and M. Halle, 1956. *Fundamentals of language.* Gravenhage: Mouton & Co..

Jansche, Martin. 1999. The Mandarin third tone sandhi across dialects. The Ohio State University, ms.

Kiriloff, C. 1969. On the Auditory perception of Tones in Mandarin. *Phonetica* 20: 63-67.

Kohler K. 1990. Segmental reduction in connected speech: Phonological facts and phonetic explanations. In W.J. Hardcastle & A. Marchal (eds.) *Speech Production and Speech Modeling* (pp. 69-72). Dordrecht: Kluwer Academic Publishers.

Kratochvil, Paul, 1968. *The Chinese Language Today: features of an emerging standard.* London: Hutchinson.

Lee, Y-S, Vakoch, D.A., & L. H. Wurm, 1996. Tone Perception in Cantonese and Mandarin: A Cross-Linguistic Comparison. *Journal of Psycholinguistic Research.*

Nosofsky, R. M. 1992. Similarity scaling and cognitive process models. *Annual Review of Psychology*, 1992, 43, 25-53.

Shen, X. S. & Lin, M. 1991. A perceptual study of Mandarin Tone 2 and 3. *Language and Speech*, 34 (2), 145-156.

Shepard, R. N. The circumplex and related topological manifolds in the study of perception. In Shye, S. (Ed.), *Theory construction and data analysis in the social sciences*. San Francisco: Jossey-Bass, 1978.

Shepard, R. N., Kilpatric, D. W., & J. P. Cunningham, 1975. The internal representation of numbers. *Cognitive Psychology*, 1975, 7, 82-138.

Shih, Chi-lin, 1997. Mandarin third tone sandhi and prosodic structure. In Wang Jialing and Norval Smith (eds.) *Studies in Chinese Phonology*. Berlin & New York: Mouton de Gruyter, 1997.

Steriade, D. 2001. A perceptual account of directional asymmetries in assimilation and cluster reduction. In E. Hume & K. Johnson (eds.) *The Role of Speech Perception in Phonology* (pp 219-250). New York: Academic Press.

Takane Y. & J. Sergent, 1983. Multidimensional Scaling Models for Reaction Times and Same-different Judgments. *Psychometrika*, 1983, Vol. 48, No. 3, 393-423.

Trubetzkoy, N. S. 1969. *Principles of phonology*. C. Baltaxe (translator). Berkeley and Los Angeles: University of California Press, 1969.

Yip, M. 1980. The Tonal Phonology of Chinese. Ph.D. dissertation. MIT. Distributed by the Indiana University Linguistics Club.

Zhang, Zheng-sheng, 1988. Tone and Tone Sandhi in Chinese. OSU Ph.D. dissertation.

Appendix I – Multidimensional Scaling of the median data

Table 6. Median RT values for correct responses (in milliseconds)

tonal pairs	T1/T2		T1/T3		T1/T4	
	T1T2	T2T1	T1T3	T3T1	T1T4	T4T1
Chinese	546	509	536.5	540	551	554
English	504	621	493	520	575.5	566
tonal pairs	T2/T3		T2/T4		T3/T4	
	T2T3	T3T2	T2T4	T4T2	T3T4	T4T3
Chinese	**656**	**675**	484	559	502	519
English	**698**	**627**	587	547	583	590

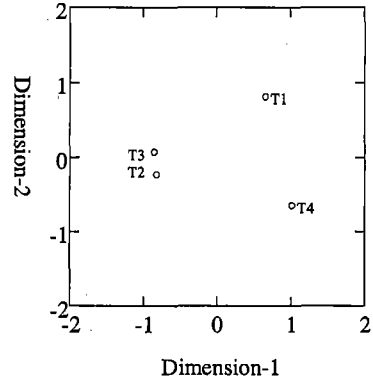

Figure 9. MDS analysis on median reaction time data of the Chinese listeners. [stress = 0.00, Proportion of variance (RSQ) = 1.0]

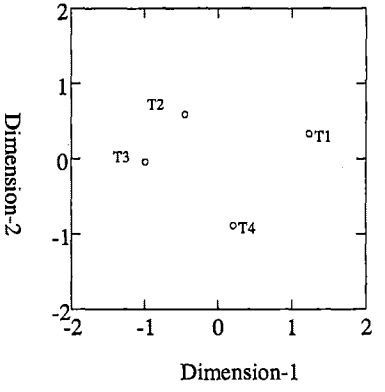

Figure 10. MDS analysis on median reaction time data of the English listeners. [stress = 0.00, Proportion of variance (RSQ) = 1.0]

Appendix II – Multidimensional Scaling of the error data

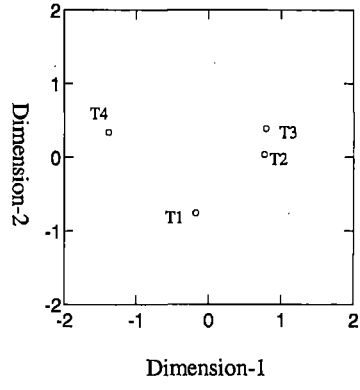

Figure 11. MDS analysis on the error data of the Chinese listeners. [stress = 0.00, Proportion of variance (RSQ) = 1.0]

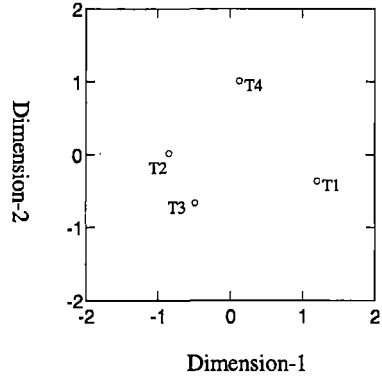

Figure 12. MDS analysis on the error data of the English listeners. [stress = 0.00, Proportion of variance (RSQ) = 1.0]